UNIVERSITÄT BIELEFELD

Decoupling Pragmatics: Discriminative Decoding for Referring Expression Generation

Simeon Schüz, Sina Zarrieß
Bielefeld University

ReInAct 2021

Scope of this work

- Task: **Referring Expression Generation (REG)**
  - Generating expressions for entities, which allow their identification in a given context
- We explore **discriminative decoding for REG**
  - Integrating pragmatic reasoning into inference
  - Shown to be effective for image captioning; not tested for REG yet
  - Shares traits with pre-neural REG approaches

Outline

1) Background
2) Decoding Methods
3) Experimental Set-Up
4) Results
5) Conclusion

Background

- Referring Expression Generation (REG):
  - Long-standing field of interest in Computational Linguistics
  - Goal: Generating expressions for entities, which allow their identification in context

d_1 d_2 d_3

d_2: "The woman"

d_1: "The man on the left" / "The man with the tie" / ...

Example: Krahmer & Van Deemter 2012

Background

Traditional (pre-neural) REG

- Entities represented in terms of pairs of **symbolic attributes and values**
- Focus on **content selection**, i.e. algorithms for attribute selection

→ **Interpretable**: Explicit algorithms, which operate on symbolic information

→ **Distinction** between semantic information (information about entities) and pragmatic processing (content selection algorithms)

d_1 d_2 d_3

Object	type	clothing	position
d_1	man	wearing suit	left
d_2	woman	wearing t-shirt	middle
d_3	man	wearing t-shirt	right

Example: Krahmer & Van Deemter 2012

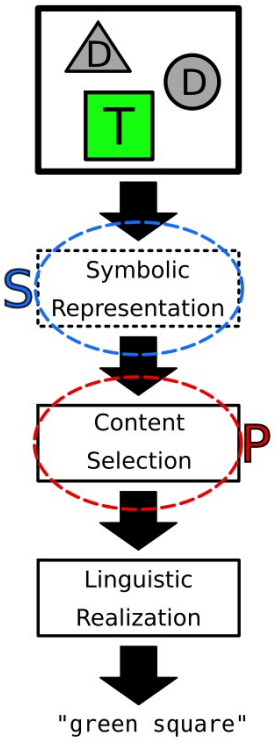
Background

Traditional (pre-neural) REG

- Entities represented in terms of pairs of **symbolic attributes and values**
- Focus on **content selection**, i.e. algorithms for attribute selection

→ **Interpretable**: Explicit algorithms, which operate on symbolic information

→ **Distinction** between semantic information (information about entities) and pragmatic processing (content selection algorithms)



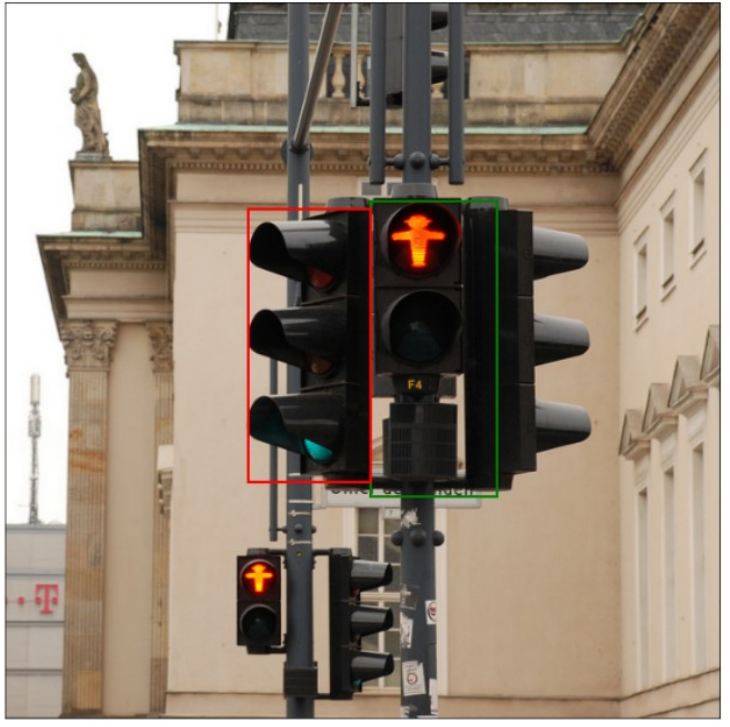
Background

Neural REG

- More realistic settings possible
- Comprises several aspects of the REG task

→ **Less interpretable**: Continuous representations & black box models

→ **Distinction** between semantic and pragmatic processing is **less clear**



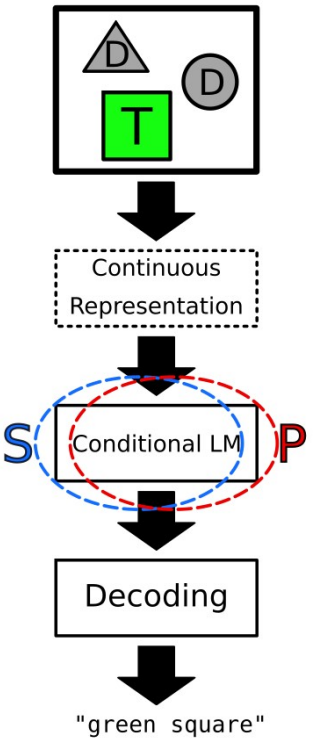
Background

Neural REG

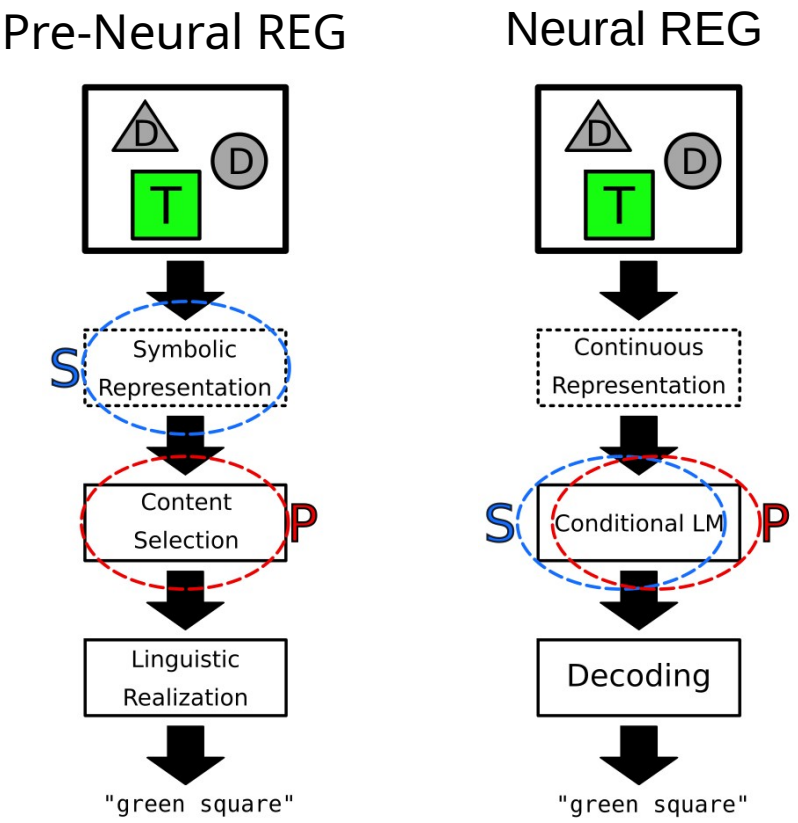
- More realistic settings possible
- Comprises several aspects of the REG task

→ **Less interpretable**: Continuous representations & black box models

→ **Distinction** between semantic and pragmatic processing is **less clear**



Background


Pre-Neural REG
D
D
T
S
Symbolic
Representation
Content
Selection
P
Linguistic
Realization
"green square"
Neural REG
D
D
T
Continuous
Representation
S
Conditional LM
P
Decoding
"green square"

Background

Discriminative Decoding in Image Captioning

- similar objective as in REG: distinguish target from distractor images
- Pragmatic reasoning during decoding: Boost informative & inhibit ambiguous word tokens at each time step

→ **more interpretable** than neural REG: pragmatic reasonining over symbolic items (words)

→ **(more) clear distinction** between semantic and pragmatic processing (CNLM vs. decoding stage)

$Greedy$	a clock tower with a clock on top of it
$Nucleus_{p0.7-t1.0}$	a building that has a clock tower in the center
$ES-Beam_{\lambda 0.5}$	a tall clock tower with trees in the background
$RSA-Beam_{\alpha 1.0}$	a view of a tall clock tower with trees in the background

Schüz et al. 2021

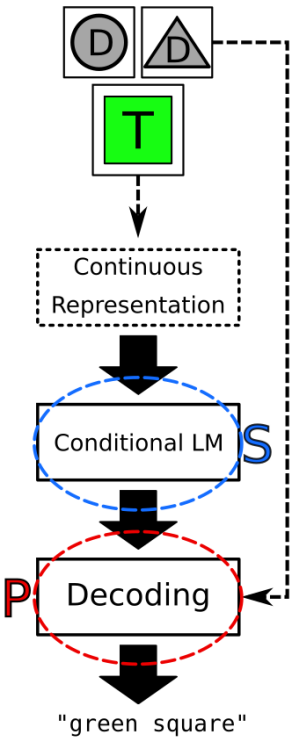
Background

Discriminative Decoding in Image Captioning

- similar objective as in REG: distinguish target from distractor images
- Pragmatic reasoning during decoding: Boost informative & inhibit ambiguous word tokens at each time step

→ **more interpretable** than neural REG: pragmatic reasonining over symbolic items (words)

→ **(more) clear distinction** between semantic and pragmatic processing (CNLM vs. decoding stage)

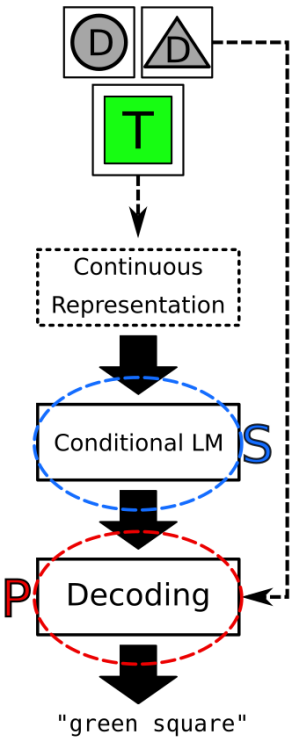


Background

Discriminative Decoding in Image Captioning

- Investigated in Schüz et al. 2021:
  - Decreases **likelihood** to ground-truth annotations
  - Increases pragmatic **informativity**
  - Increases (global) **linguistic diversity**

Do we see the same patterns in neural REG?



Background

Discriminative Decoding in neural REG

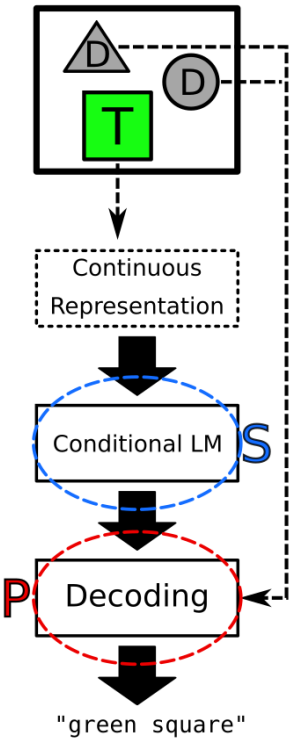
→ topic in this work

- more natural setting than discriminative image captioning
- “middle ground” between pre-neural and neural REG

→ builds upon neural models

→ more interpretable than neural REG

→ (conceptually) clear distinction between semantic and pragmatic processing (“literal” generation model vs. decoding stage)



Decoding Methods

- We compare four decoding methods:

1) Focus on **Likelihood**
    1) Greedy search
    2) Beam search
2) Focus on **Informativity**
    1) RSA
    2) Emitter / Supressor

Decoding Methods

Decoding for Likelihood:

1) Greedy Search

- Pick the most probable token at each step

2) Beam Search

- Expand multiple hypotheses simultaneously

Decoding Methods

Decoding for Informativity:

1) **RSA** (Cohn-Gordon et al. 2018)
2) **Emitter / Supressor** (Vedantam et al. 2017)

- Both methods are conceptually similar:
  - Reasoning at each time step during decoding
  - Reweighting predictions of a context-agnostic generation model
- In both cases:
  - Reasoning procedure embedded in a beam search scheme

Decoding Methods

- RSA Decoding (Cohn-Gordon et al. 2018)
  - Based on the *Rational Speech Acts* Framework
  - Integrates pragmatic reasoning into the iterative unrolling of the generation model: How would a listener interpret the utterance?

Decoding Methods

- *Literal Speaker (S_0):* Generation Model; assigns initial probabilities to token candidates
- *Pragmatic Listener (L_0):* Assesses whether token candidates allow identifying the target
- *Pragmatic Speaker (S_0):* Chooses the most informative token candidates
- α : Rationality Parameter

$$L_0(w|u) \propto \frac{S_0(u|w) * P(w)}{\sum_{w' \in W} S_0(u|w') * P(w')}$$

$$S_1(u|w) \propto \frac{L_0(w|u)^{\alpha} * P(u)}{\sum_{u' \in U} L_0(w|u')^{\alpha} * P(u')}$$

Decoding Methods

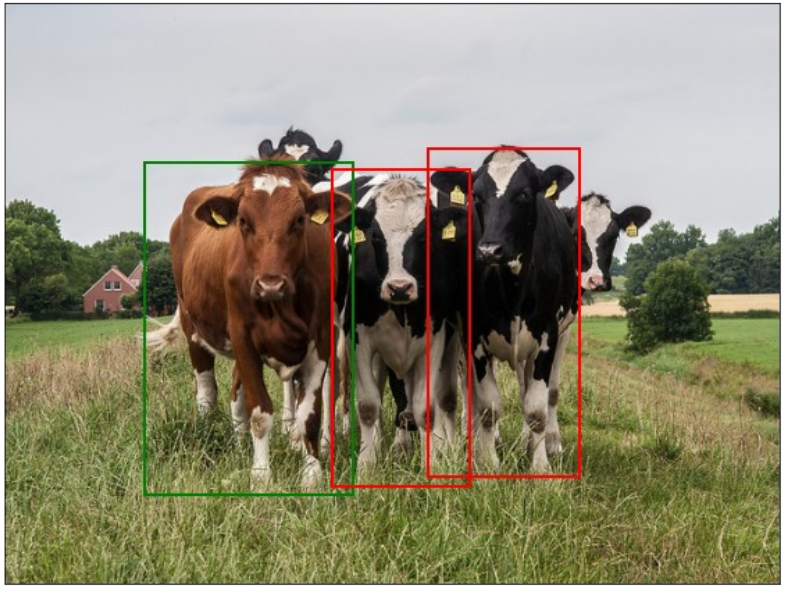
- Emitter / Supressor (Vedantam et al. 2017)
  - Base model as *Emitter*
  - *Supressor* function rates informativity
- Extended version from Schüz et al. (2021) for multiple distractors
- λ: Rationality Parameter

$$\Delta(I_t, D) = \arg\max_s \sum_{\tau=1}^{T} \sum_{i=1}^{|D|} \log \frac{p(s_\tau | s_{1:\tau-1}, I_t)}{p(s_\tau | s_{1:\tau-1}, D_i)^{1-\lambda}}$$

Experimental Set-Up / Data

RefCOCO, **RefCOCO+** (Kazemzadeh et al., 2014), **RefCOCOg** (Mao et al., 2016)

- English referring expressions & bounding boxes for objects in images from MSCOCO (Lin et al., 2014)



Example from RefCOCO.
Annotations:

- brown cow
- brown
- red one

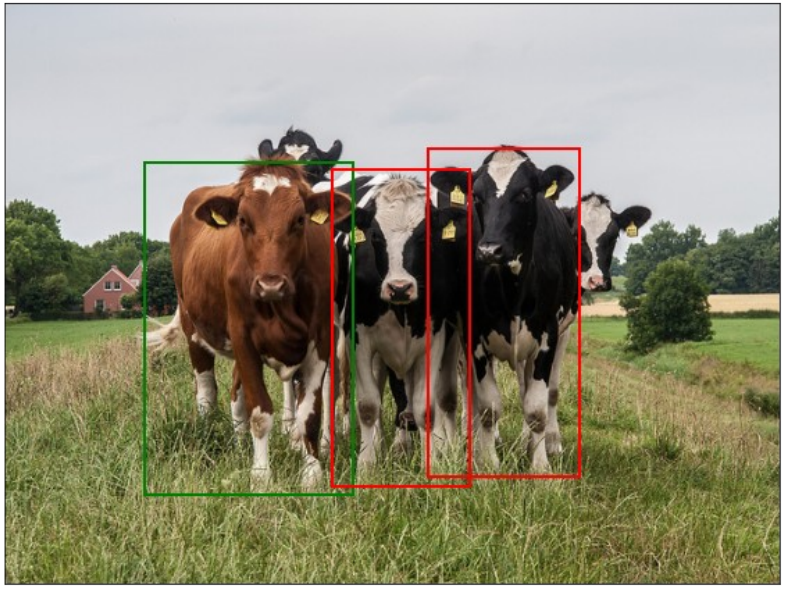
Experimental Set-Up / Data

- Data Filtering:
  - Removed images with single object annotations
- Test splits:
  - Multiple test sets for RefCOCO/RefCOCO+: References to humans (*testA*) or other objects (*testB*)
  - Number of objects:

    ~ 1950 testA / testA+

    ~ 1750 testB / testB+

    ~ 4000 testg



Example from RefCOCO.
Annotations:

- brown cow
- brown
- red one

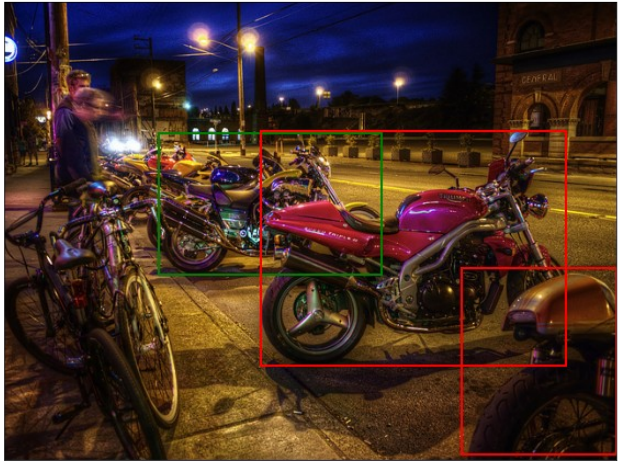
Experimental Set-Up / Model

- Adaptive Attention Model (Lu et al. 2017)
  - Used for all decoding methods
  - Adapted for REG:
    - Visual input: Object bounding boxes
    - 7 location features encoded jointly with visual input

Experimental Set-Up / Evaluation

- Likelihood
  - BLEU1, CIDEr
- Diversity:
  - TTR1, TTR2, % coverage
- Informativity
  - Detection accuracy from external Referring Expression Comprehension model (Luo et al. 2020)
  - Model proposes bounding boxes for described objects
  - classified true / false based on IOU with ground-truth annotations
    - true if IOU > 0.5

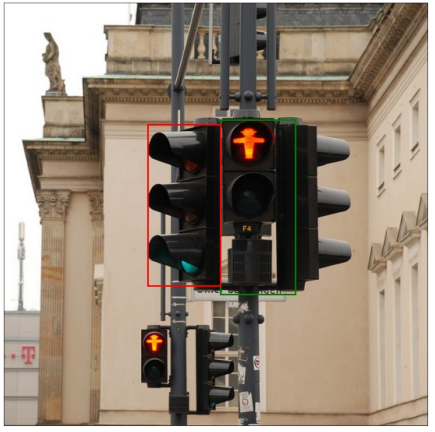
Results / Qualitative Examples



Dataset: RefCOCO+
Annotations:
green motorcycle
bike behind purple one
bike behind red bike

Greedy	red bike
Beam	red bike
$ES_{\lambda 0.7}$	red bike
$ES_{\lambda 0.5}$	blue bike
$ES_{\lambda 0.3}$	blue bike
$RSA_{\alpha 0.5}$	blue bike
$RSA_{\alpha 1.0}$	blue bike
$RSA_{\alpha 5.0}$	blue bike

Results / Qualitative Examples



Greedy traffic light
Beam traffic light

$ES_{\lambda 0.5}$ red light
$RSA_{\alpha 5.0}$ stop light

Beam right dog
Greedy right dog

$ES_{\lambda 0.5}$ right cow
$RSA_{\alpha 5.0}$ far right

Results / Likelihood

- Discriminative Decoding **decreases** likelihood
- Generation model was trained on likelihood objectives; ES & RSA deviate from model predictions

	testA			testB			testA+			testB+			testg		
	BL_1	CDr		BL_1	CDr		BL_1	CDr		BL_1	CDr		BL_1	CDr	
greedy	**53.2**	0.79		**56.2**	1.27		**45.6**	**0.68**		**33.9**	0.64		45.0	0.71	
beam	52.8	**0.81**		55.4	**1.31**		40.4	0.66		32.5	**0.75**		**45.0**	**0.79**	
$ES_{\lambda 0.7}$	52.9	0.80		55.7	1.24		40.3	0.67		30.9	0.71		44.1	0.74	
$ES_{\lambda 0.5}$	49.8	0.73		53.3	1.11		37.1	0.60		27.4	0.61		40.4	0.63	
$ES_{\lambda 0.3}$	35.8	0.53		44.5	0.81		23.1	0.37		21.0	0.40		29.5	0.37	
$RSA_{\alpha 0.5}$	50.2	0.77		55.1	1.26		32.7	0.58		28.2	0.67		43.5	0.70	
$RSA_{\alpha 1.0}$	50.4	0.77		54.8	1.22		33.3	0.59		27.5	0.65		43.0	0.69	
$RSA_{\alpha 5.0}$	50.9	0.75		52.7	1.05		35.4	0.57		25.7	0.58		40.2	0.61	
human	-	-		-	-		-	-		-	-		-	-	

Results / Diversity

- Discriminative Decoding **increases** diversity
- Diverging from the model predictions leads to more variation and the usage of a larger vocabulary

	testA			testB			testA+			testB+			testg		
	T_1	T_2	cov.	T_1	T_2	cov.	T_1	T_2	cov.	T_1	T_2	cov.	T_1	T_2	cov.
greedy	7.0	29.2	4.1	10.6	46.5	5.6	8.7	28.2	4.2	16.5	48.1	7.7	16.7	40.0	10.7
beam	6.6	27.8	3.6	10.3	48.6	5.2	9.4	32.7	4.1	16.4	56.4	6.9	17.5	42.4	10.4
$ES_{\lambda 0.7}$	8.3	35.0	4.8	11.6	49.7	6.1	12.5	42.6	5.4	19.3	64.0	7.9	19.8	48.0	13.0
$ES_{\lambda 0.5}$	11.6	46.4	6.9	13.4	54.4	7.6	16.1	54.4	7.9	22.2	70.3	9.5	22.7	56.4	16.5
$ES_{\lambda 0.3}$	**17.5**	**60.7**	**12.6**	**18.4**	**62.6**	**13.0**	**22.7**	**65.8**	**13.2**	**29.2**	**80.3**	**14.8**	**28.7**	**71.0**	**23.6**
$RSA_{\alpha 0.5}$	7.8	33.7	4.3	11.6	50.2	5.9	12.1	43.2	5.0	18.5	61.0	7.4	19.8	48.1	12.6
$RSA_{\alpha 1.0}$	7.8	34.6	4.4	11.8	51.6	5.9	13.0	47.0	5.7	19.7	64.7	7.9	20.4	49.8	13.6
$RSA_{\alpha 5.0}$	9.4	39.0	5.7	13.5	55.8	7.2	16.1	54.1	7.6	23.9	74.4	10.6	22.5	57.7	16.3
human	24.9	71.7	22.4	27.6	79.6	23.1	31.2	80.6	20.9	39.6	91.2	26.2	34.0	77.8	44.4

Results / Informativity

- Discriminative Decoding **increases** detection accuracy
- Gain is rather modest (highest gain is ~10 % on testB+)

	testA			testB			testA+			testB+			testg		
	[illegible]	[illegible]	det.	[illegible]	[illegible]	det.	[illegible]	[illegible]	det.	[illegible]	[illegible]	det.	[illegible]	[illegible]	det.
greedy	[illegible]	[illegible]	75.1	[illegible]	[illegible]	66.4	[illegible]	[illegible]	63.8	[illegible]	[illegible]	43.0	[illegible]	[illegible]	54.4
beam	[illegible]	[illegible]	75.0	[illegible]	[illegible]	66.0	[illegible]	[illegible]	62.4	[illegible]	[illegible]	43.6	[illegible]	[illegible]	54.8
$ES_{\lambda 0.7}$	[illegible]	[illegible]	77.6	[illegible]	[illegible]	70.5	[illegible]	[illegible]	67.1	[illegible]	[illegible]	46.9	[illegible]	[illegible]	57.1
$ES_{\lambda 0.5}$	[illegible]	[illegible]	**80.6**	[illegible]	[illegible]	**71.8**	[illegible]	[illegible]	**68.7**	[illegible]	[illegible]	47.3	[illegible]	[illegible]	57.0
$ES_{\lambda 0.3}$	[illegible]	[illegible]	79.1	[illegible]	[illegible]	71.3	[illegible]	[illegible]	66.5	[illegible]	[illegible]	**48.1**	[illegible]	[illegible]	54.6
$RSA_{\alpha 0.5}$	[illegible]	[illegible]	76.0	[illegible]	[illegible]	69.4	[illegible]	[illegible]	62.8	[illegible]	[illegible]	44.4	[illegible]	[illegible]	55.7
$RSA_{\alpha 1.0}$	[illegible]	[illegible]	76.4	[illegible]	[illegible]	69.1	[illegible]	[illegible]	63.2	[illegible]	[illegible]	44.6	[illegible]	[illegible]	56.1
$RSA_{\alpha 5.0}$	[illegible]	[illegible]	79.3	[illegible]	[illegible]	70.9	[illegible]	[illegible]	66.3	[illegible]	[illegible]	46.8	[illegible]	[illegible]	**57.4**
human	-	-	84.4	-	-	74.2	-	-	72.6	-	-	57.9	-	-	63.7

Results / Informativity

Why is there no higher increase?

1. The comprehension task is **difficult**:

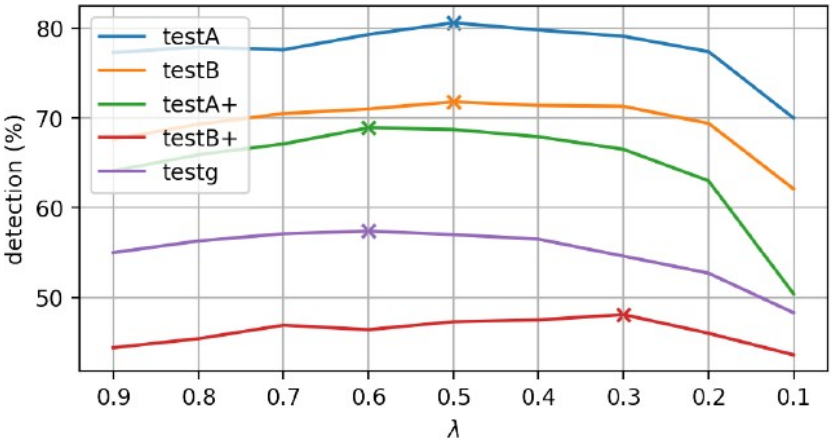
- For some parts of our data, the evaluation model struggles even with human annotations (e.g. testB+)
- “achievable” results are thus limited

2. **Literal expressions** are already quite **informative**:

- Generation model is trained on informative expressions (unlike image captioning) and might catch up effective patterns (color, location, …)
- This might render additional layers of reasoning less effective

Results / Informativity

- Lower detection results for higher rationality
- Very high rationalities generate less well-formed expressions & deviate from the training data of the evaluation model



Conclusion

- Discriminative decoding: Appealing, since it combines traits from traditional & neural REG
- Similar results as for Image Captioning:
  - Decreased likelihood to human annotations; increased diversity & pragmatic informativity
- Margin of gain in informativity is smaller than expected
  - Hypothesis: Literal REG model implicitly learns to refer informatively by informative patterns in the data
  - Especially significant in our case; but relevant for REG research in general

Citations

- Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443.
- Emiel Krahmer and Kees van Deemter. 2012. Computational Generation of Referring Expressions: A Survey. In *Computational Linguistics (38, 1),* pages 173–218.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.
- Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Simeon Schüz, Ting Han and Sina Zarrieß. 2021. Diversity as a By-Product: Goal-oriented Language Generation Leads to Linguistic Variation. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422.
- Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.